Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models

Erik Kjellström[1],*, Fredrik Boberg[2], Manuel Castro[3], Jens Hesselbjerg Christensen[2], Grigory Nikulin[1], Enrique Sánchez[3]

[1]Rossby Centre, Swedish Meteorological and Hydrological Institute, 60176 Norrköping, Sweden
[2]Danish Climate Centre, Danish Meteorological Institute, Lyngbyvej 100, 2100 Copenhagen Ø, Denmark
[3]Facultad de Ciencias del Medio Ambiente, Universidad de Castilla-La Mancha (UCLM), Toledo, Spain

ABSTRACT: We evaluated daily and monthly statistics of maximum and minimum temperatures and precipitation in an ensemble of 16 regional climate models (RCMs) forced by boundary conditions from reanalysis data for 1961–1990. A high-resolution gridded observational data set for land areas in Europe was used. Skill scores were calculated based on the match of simulated and observed empirical probability density functions. The evaluation for different variables, seasons and regions showed that some models were better/worse than others in an overall sense. It also showed that no model that was best/worst in all variables, seasons or regions. Biases in daily precipitation were most pronounced in the wettest part of the probability distribution where the RCMs tended to overestimate precipitation compared to observations. We also applied the skill scores as weights used to calculate weighted ensemble means of the variables. We found that weighted ensemble means were slightly better in comparison to observations than corresponding unweighted ensemble means for most seasons, regions and variables. A number of sensitivity tests showed that the weights were highly sensitive to the choice of skill score metric and data sets involved in the comparison.

KEY WORDS: Regional climate models · Probability distributions · Skill score metrics · Weighted ensemble · Europe

1. INTRODUCTION

Regional climate models (RCMs) are powerful tools that have the potential to provide high-resolution information both in space and time on the properties of a range of climate variables. On some occasions, data from a range of different RCM simulations are considered. A question in this context is how individual regional climate change scenarios can be converted into probabilistic climate change signals. This is relatively straightforward if models are used as they are without acknowledging the fact that different models are more or less good at representing today's climate. However, if model performance is to be considered, the situation becomes more complicated. To take this larger complexity into account, one idea, pursued in the ENSEMBLES pro-

ject (Hewitt & Griggs 2005, van der Linden & Mitchell 2009), is the weighting of RCMs according to how adequately they simulate the recent past climate of the last decades. In this context, a number of different weights measuring various aspects of RCM realism have been considered (Christensen et al. 2010, this Special). Here, as part of the ENSEMBLES effort, we investigated the match of RCM-simulated probability density distributions for maximum and minimum temperature and precipitation to those based on observational data.

Forced with boundary data from reanalysis products such as ERA40 (Uppala et al. 2005), RCMs have been shown to adequately simulate important aspects of the climate of the recent past decades, including not just long-term seasonal means but also extreme events of e.g. temperature and precipitation (Christensen et al.

2007). When forced by global climate models (GCMs), RCMs have been found to sometimes produce large biases in the mean climate (e.g. Räisänen et al. 2004, Jacob et al. 2007) and even more so in terms of extreme conditions (e.g. Moberg & Jones 2004, Beniston et al. 2007, Kjellström et al. 2007, Rockel & Woth 2007, Nikulin et al. 2010). Kjellström et al. (2007) showed that there is a considerably larger spread between RCMs at the tails of the temperature probability distributions than in simulated means both in today's climate and in the simulated climate change signals. This spread in the climate change signal was large even when the forcing conditions were taken from the same GCM following the same emissions scenario. Such a large spread indicates a sensitivity of the results to the formulation of the RCM.

Kjellström et al. (2007) investigated the daily maximum and minimum temperature by studying fixed percentiles from the empirical probability distributions. This approach gave a first indication of some of the quantiles of the probability distributions which had the largest biases, but did not give information on the whole distribution. Here we evaluated the models' ability to represent the whole probability distributions simultaneously by using the skill score (SS) metric presented by Perkins et al. (2007). Earlier, Boberg et al. (2009a,b) used this method to compare the daily precipitation in the PRUDENCE (Christensen & Christensen 2007) and ENSEMBLES simulations by comparison with the European Climate Assessment (ECA; Klein Tank et al. 2002) observational data set. In addition to evaluating daily data of precipitation and minimum and maximum temperatures, we also performed a comparison of probability distributions of monthly precipitation data. By looking at those 2 data sets simultaneously, we obtained a broader picture of how RCMs reproduce the observed climate. Examples are given of model behaviour in different regions of Europe and in different seasons. Results from 16 downscaling experiments of the ERA40 reanalysis data to 25 km horizontal resolution were evaluated against observations. Further, we calculated combined weights based on the evaluation for each model. The resulting weights (SSs) are dimensionless numbers ranging between 0 and 1 that are suitable for weighting model results in order to construct probabilistic climate change scenarios based on a given experimental setup as in the EN-

SEMBLES project (Christensen et al. 2010, Déqué & Somoto 2010, both this Special). Finally, we investigated whether a weighted mean outperforms an unweighted mean when evaluated against observations.

2. DATA AND METHODS

2.1. RCM data

We used data from 16 climate simulations with RCMs run at approximately 25 km grid spacing (Table 1). All models that operated on a rotated latitude–longitude grid had the majority of their model domain in common (Fig. 1), i.e. they worked on an identical grid (see Table 1). The other models operated on Lambert-conformal grids that were not internally identical and not identical to the rotated ones. Data from these other models were interpolated to the common rotated latitude–longitude grid shared by the others. All models were forced by lateral boundary conditions and sea surface temperatures from the European Centre for Medium Range Weather Forecasts (ECMWF) reanalysis product ERA40 (Uppala et al. 2005).

From the models, we used time series of daily minimum and maximum temperatures at the 2 m level and daily precipitation as well as monthly precipitation for

Table 1. Regional climate models from which data were analysed. RPN: Recherche en Prévision Numérique. Full forms of other institutes in Table 1 of Christensen et al. (2010, this Special)

Number	Institute	Model	Time step (min)	Source
1[a]	C4I	RCA3	15	Kjellström et al. (2005)
2	CHMI	ALADIN	15	Farda et al. (2010)
3	CNRM	RM4.5	22.5 / 6 h[b]	Radu et al. (2008)
4[a]	DMI	HIRHAM5	10	Christensen et al. (2006)
5[a]	ETH	CLM	2 / 3 h[b]	Böhm et al. (2006)
6	ICTP	RegCM3	1.25 / 10[b]	Giorgi & Mearns (1999)
7[a]	KNMI	RACMO2	12	van Meijgard et al. (2008)
8	Met.No	HIRHAM	3.75	Haugen & Haakenstad (2006)
9[a]	Hadley Centre	HadRM3Q0	5	Collins et al. (2009)
10[a]	Hadley Centre	HadRM3Q3	5	Collins et al. (2009)
11[a]	Hadley Centre	HadRM3Q16	5	Collins et al. (2009)
12[a]	MPI-M	REMO	2	Jacob (2001)
13	OURANOS	MRCC4.2.3	10	Plummer et al. (2006)
14[a]	RPN	GEMLAM	12	Côté et al. (1998), Zadra et al. (2008)
15[a]	SMHI	RCA3.0	15	Kjellström et al. (2005)
16	UCLM	PROMES	0.83 / 10[b]	Sanchez et al. (2004)

[a]Models that used the same rotated longitude–latitude grid
[b]The first number (left) represents the RCM time step; the second number (right) is the sampling interval used for maximum and minimum temperatures
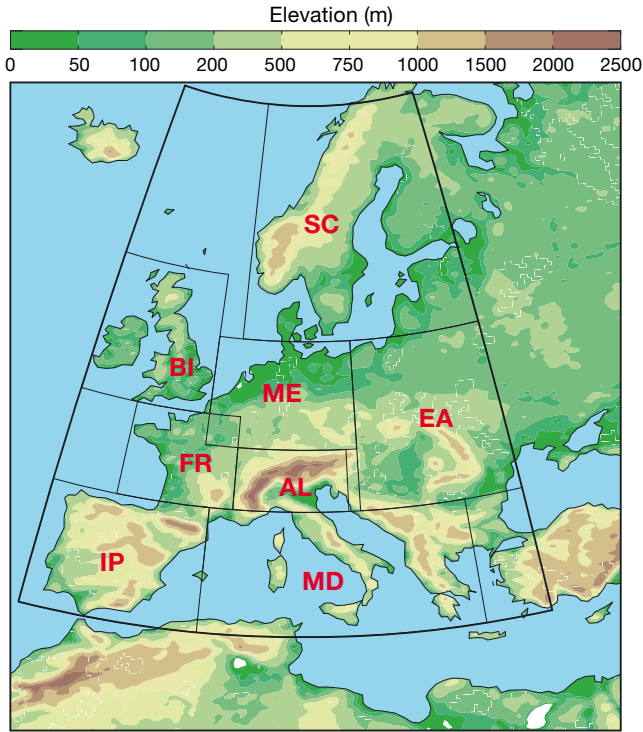
Fig. 1. Domains of the regional climate models (RCMs) running on the same rotated latitude–longitude grid. The colours depict altitude (from the SMHI-RCA3.0 model). Indicated are also the 8 sub-domains (BI: British Isles, IP: Iberian Peninsula, FR: France, ME: mid-Europe, SC: Scandinavia, MD: Mediterranean region, AL: Alps, EA: eastern Europe) and the larger European domain (EU) for which SSs were calculated

the time period 1961–1990. Data were downloaded from the ENSEMBLES regional data distribution centre at DMI (http://ensemblesrt3.dmi.dk). The time steps used in the different RCMs for calculating daily maximum and minimum temperatures varied between 2 and 15 min, except for the ETH and CNRM models, for which they were estimated based on instantaneous 3- or 6-hourly data (Table 1).

2.2. Daily and monthly observational data

Model results were compared to daily minimum and maximum temperature and precipitation from a gridded observational data set (E-OBS) that is based on the largest existing pan-European data set with daily data extending back to 1950 (Haylock et al. 2008, Klok & Klein Tank 2009). A benefit from using this particular data set is that it is constructed on the same rotated latitude/longitude grid that was used by most of the RCMs, implying that no further interpolation was needed for most of the RCMs. In a comparison with other existing data sets from more dense networks in smaller regions,

Hofstra et al. (2009) found that E-OBS correlated well with these other data sets. However, they also found that relative differences in precipitation could be large, and that they were usually biased toward lower values in E-OBS. Monthly precipitation was calculated as the sum of all precipitation on each day of a month.

2.3. Metrics used for daily data

A description of the metric we used for comparing daily minimum and maximum temperatures and precipitation to E-OBS is given by Perkins et al. (2007). Empirical probability distribution functions (PDFs) were first constructed by binning data in N number of bins according to temperature or precipitation amount, and then generating a dimensionless 'match metric' or SS based on the overlap of the RCM and observation PDFs. A perfect overlap results in an SS of 1, whereas the score is close to 0 for a low degree of overlap:

$$SS = \sum_{bin=1}^{bin=N} \min(PDF_{RCM}, PDF_{E\text{-}OBS}) \quad (1)$$

2.4. Metrics used for monthly precipitation data

For the monthly precipitation data, we used the same SS as defined in Eq. (1) but for the discussion also an additional metric to test the sensitivity of the resulting weights to the choice of scoring metric. This second metric was taken from Sánchez et al. (2009) and consisted of a collection of 5 functions measuring different aspects of model probability distribution characteristics.

$$f_1 = 1 - \left(\frac{|A_{RCM} - A_{CRU}|}{2 \cdot A_{CRU}} \right)^{0.5} \quad (2)$$

$$f_2 = 1 - \left(\frac{|A_{RCM}^+ - A_{CRU}^+|}{2 \cdot A_{CRU}^+} \right)^{0.5} \quad (3)$$

$$f_3 = 1 - \left(\frac{|A_{RCM}^- - A_{CRU}^-|}{2 \cdot A_{CRU}^-} \right)^{0.5} \quad (4)$$

$$f_4 = 1 - \left(\frac{|\overline{P_{RCM}} - \overline{P_{CRU}}|}{2 \cdot \overline{P_{CRU}}} \right)^{0.5} \quad (5)$$

$$f_5 = 1 - \left(\frac{|\sigma_{RCM} - \sigma_{CRU}|}{2 \cdot \sigma_{CRU}} \right)^{0.5} \quad (6)$$

where A_{RCM} and A_{CRU} are the areas below the empirical cumulative distribution functions (CDFs obtained from normalised PDFs) of the RCMs and observations respectively, A^+ and A^- are the corresponding areas to the right and left of the 50th percentile, respectively, P is the spatial and temporal average of precipitation in each season and region and σ is the standard devia-

tion. Each of the factors takes into account different aspects of the behaviour of the model precipitation. These are: the distribution as a whole in terms of mean (Eq. 5) and total area (Eq. 2), the precipitation amounts for more intense precipitation in the upper (Eq. 3) and moderate precipitation in the lower (Eq. 4) half of the distribution divided by the median and the width of the distribution through the variance (Eq. 6). These 5 measures are all constructed so that a value close to 1 indicates that the RCM is close to observations while a value close to 0 indicates that the RCM is far from the observations. Finally, the 5 numbers are multiplied by each other giving a final dimensionless SS that also is between 0 and 1.

$$SS = f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \tag{7}$$

2.5. Data handling including area averaging

SSs were calculated for 8 European regions (Fig. 1) as well as for the entire continental land area between 36° and 70° N and 10.5° W and 30° E (EU) for each nominal season (winter: DJF, spring: MAM, summer: JJA, autumn: SON) and as an annual mean calculated as the average of the SSs for the 4 seasons (annual: ANN).

Daily maximum and minimum temperatures were binned into 0.5°C intervals for all grid boxes for which observational data were available, i.e. land areas. Different methods of averaging over different regions were tested for all models: (1) by pooling data from all grid points within each region into 1 common probability distribution before calculating the SSs, and (2) by first averaging temperatures for each day in a region and then calculating the SSs. We note here that the first method keeps all fine details but mixes spatial and temporal variability. The second method smoothes local, fine details. We also used a third method (3), in which SSs were first calculated for each grid box, and thereafter averaged over the area. In this case, data from models operating on other grids were first interpolated to the common rotated latitude–longitude grid before SSs are calculated. The first method was our reference, while the 2 others were only used to test the sensitivity of the score to the aggregation procedure.

For the daily precipitation analyses, we defined dry days as days with precipitation <1 mm and removed these from the precipitation data. PDFs were then calculated from the remaining daily data for each region and model by first binning the data into bins of 1 mm width starting at 1 mm. The impact of using a threshold is discussed in Section 4.3.

For the monthly precipitation analysis, data were aggregated in the regions according to the first method outlined above (this section) for daily data. This pooling of data provided a large sample for deriving the empirical percentiles. We note here that these monthly data also included dry days, which were excluded from the daily precipitation analyses.

2.6. Comparing RCM results to observations at different parts of the PDFs

To obtain a more detailed picture of why some RCMs achieved high and others low SSs, we calculated biases for each RCM with regard to the observations at a number of percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99) based on all data in a region. These biases in different parts of the PDFs allowed us to identify where different RCMs had problems and may help explain why SSs differed between models.

2.7. Combination of SSs into one weight per RCM

A combined weight (w_{irs}) for each model (i), region (r) and season (s) based on our model evaluation was calculated according to:

$$w_{irs} = \frac{1}{3}\left(SS_{PRD} + (SS_{TXD} + SS_{TMD})/2 + \sqrt{SS_{PRM}}\right) \tag{8}$$

with SS as in Eq. (1). Indices PRD, PRM, TXD and TMD denote daily and monthly precipitation and daily maximum and minimum temperatures, respectively. It can be seen from Eq. (8) that the influence of monthly data was downgraded, as we took the square root of the resulting weight (SS_{PRM}) when combining it with the others. The rationale for this downgrading is that while the daily data are seen as largely representing regional information and hence form a relevant RCM metric, the monthly precipitation fields are much more strongly controlled by the driving GCM but are still worthwhile to evaluate as discussed in Section 4.1 below. The dimensionless weight (w_{irs}) was normalised and takes a number between 0 and 1.

A final overall weight for each model (w_i) was calculated by averaging all individual weights (w_{irs}) over the 4 seasons and the 8 European regions. This final weight (w_i) was intended to be used in the weighting system described by Christensen et al. (2010).

3. RESULTS

3.1. Daily maximum and minimum temperatures

Fig. 2 shows the geographical distribution of the SSs for daily minimum temperatures in winter. The SSs were generally highest in an area in western central Europe and lower in the north and in the south. In par-

ticular, many models had problems in northern Scandinavia. A few RCMs stand out in that they achieved lower SSs compared to most others in much of Europe. One of these is REMO, which had excessive problems particularly in southern Europe. This is related to an unrealistically high occurrence of temperature values around 0°C that was caused by the representation of freezing and thawing processes in the soil. In REMO, as used for this study, the soil temperature was kept at 0°C until the entire soil water within a grid box was either frozen or liquid. This is unrealistic for such a subgrid process and has been modified for newer REMO versions (D. Jacob pers. comm.). REMO also achieved a low SS for daily maximum temperature
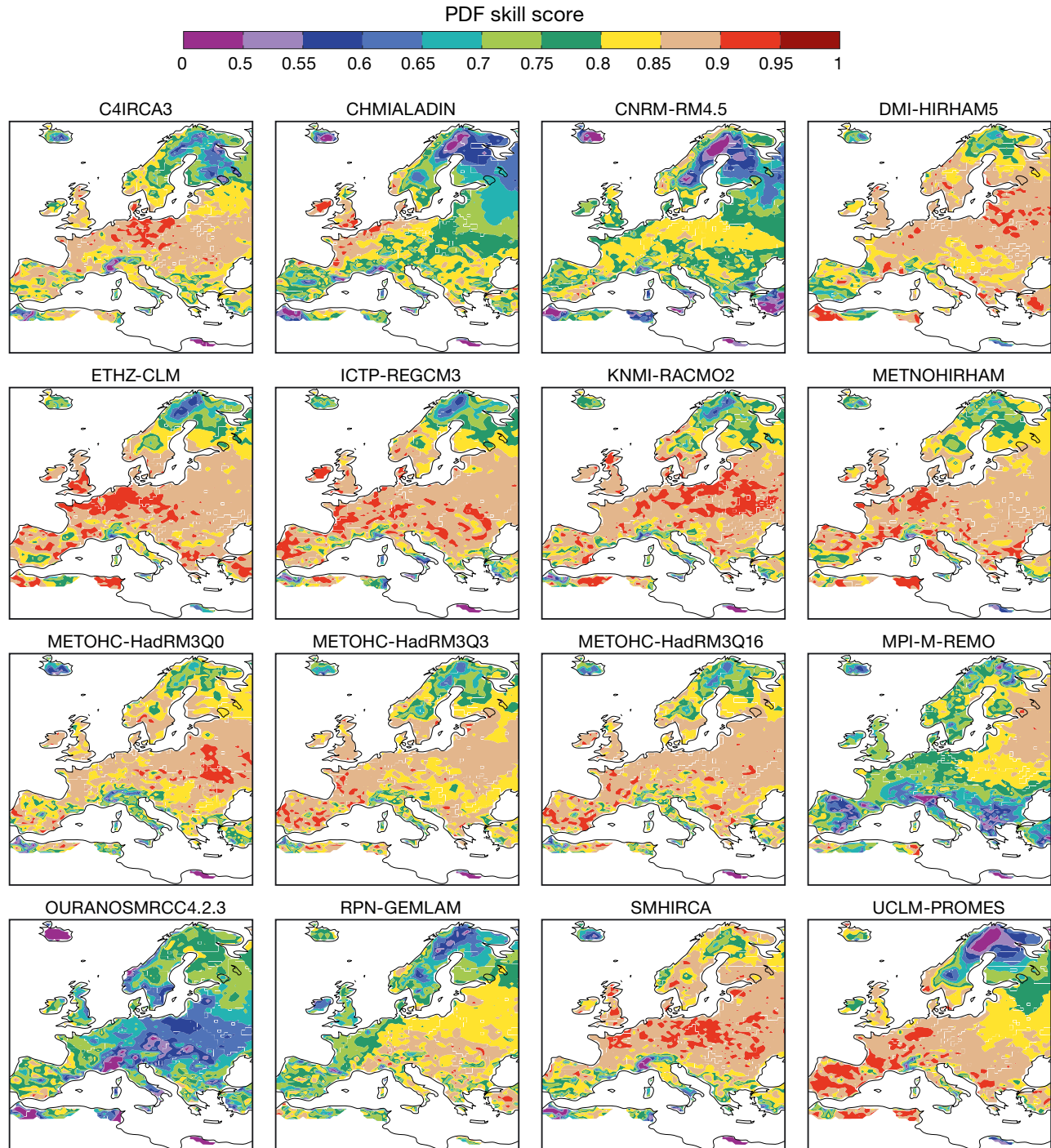

Fig. 2. Skill scores for daily minimum temperature during winter (DJF) from 1961–1990 for the 16 regional climate models (RCMs). The SS is dimensionless and takes a number between 0 and 1

during winter but in an area more to the east and north where temperatures are generally lower (not shown). MRCC4.2.3 generally had low SSs for minimum temperatures, most pronounced in parts of eastern and central Europe and also in southern Sweden. A detailed comparison for different parts of the PDFs in the Scandinavian region revealed that MRCC4.2.3 simulated temperatures were generally too low (Fig. 3), which led to an overly extensive snow cover in this region, further exacerbating the cold bias (not shown). Most other models gave higher temperatures, and differed relatively substantially from E-OBS, in northern Scandinavia (Fig. 3). A more detailed look into the biases in that region at different temperature intervals revealed that all models except REMO and MRCC4.2.3 showed the largest positive biases during the coldest days (e.g. 1st or 5th percentiles), while biases were smaller in situations with relatively warmer conditions. As the horizontal and vertical grid spacings in the RCMs were fairly coarse (25×25 km in the horizontal, and typically the lowest model level was some 50 to 100 m above ground) they did not allow for a fair representation of strong inversions or local conditions. Strong near-surface inversions are common in winter in this area. Furthermore, many of the observational stations in the area are located in valleys that may be much colder than the surrounding terrain when these strong inversions prevail. Alternatively, biases may reflect the fact that the station density is low in this area (Haylock et al. 2008), or that the stations are representative of open land conditions while model results represented a mixture of different vegetation types dominated by forests in this region
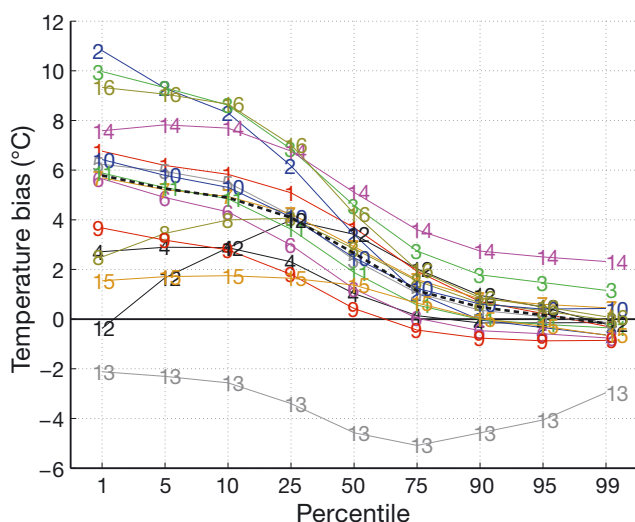

Fig. 3. For each model, the area-averaged bias over Scandinavia (SC in Fig. 1) in winter (DJF) daily minimum temperature compared to E-OBS at different percentiles. Model names given in Table 1. Dashed black line: median of the 16 models

(Nikulin et al. 2010). Taken together, these potential problems with RCMs and observational data indicate that the biases are more a consequence of the fact that the 2 data sets represent different features than an indication of a systematic model error. We also note that the problem was less pronounced for maximum temperatures in winter (not shown), further indicating that this is a problem mostly concerning the coldest situations.

Fig. 4 shows the geographical distribution of the SSs for daily maximum temperatures in summer. Some features that could be observed in many models included a generally relatively poor agreement in Scandinavia and in parts of the Mediterranean area and a 'lake problem' evident for the large lakes in northern Europe. Many RCMs that included lake models showed low SSs over the large lakes in northern Europe. The RCMs with lake models are probably more representative of the real temperatures above the lakes than the observational data set that was derived based mainly on land stations (see Kjellström et al. 2005 for a comparison of simulated and observed temperatures over Lake Ladoga and Lake Vänern). Fig. 5 reveals that the match between observations and model simulations was different in different parts of the probability distributions. A group of models including RCA3, ALADIN, RM4.5, HIRHAM5, GEMLAM and RCA3.0 exhibited increasingly large cold biases towards the high end of the probability distribution. The other RCMs showed increasingly warmer biases (or smaller cold biases in the case of HIRHAM and PROMES) towards the high end of the probability distribution. Notably, HadRM3Q3 showed very low SSs in large parts of southern Europe. These were associated with a large overestimation of daily maximum temperatures, in particular for warm occasions in this region (not shown), that also influenced the area average over Europe (Fig. 5).

3.2. Daily precipitation

Skill scores for daily precipitation were high in large parts of Europe in winter (Fig. 6). Relatively poor agreement was seen in the Mediterranean area and in the northernmost part of the model domain, including parts of Scandinavia, Finland, Russia and Iceland in most models. Compared to maximum and minimum temperatures as discussed above, the SSs for precipitation were generally higher. The relatively high SSs for precipitation were a result of good agreement in the low and central parts of the probability distributions (Fig. 7). Contrastingly, Fig. 7 reveals poor agreement to observations and a large spread among the RCMs for the wettest quartile. It also shows that the overestima-
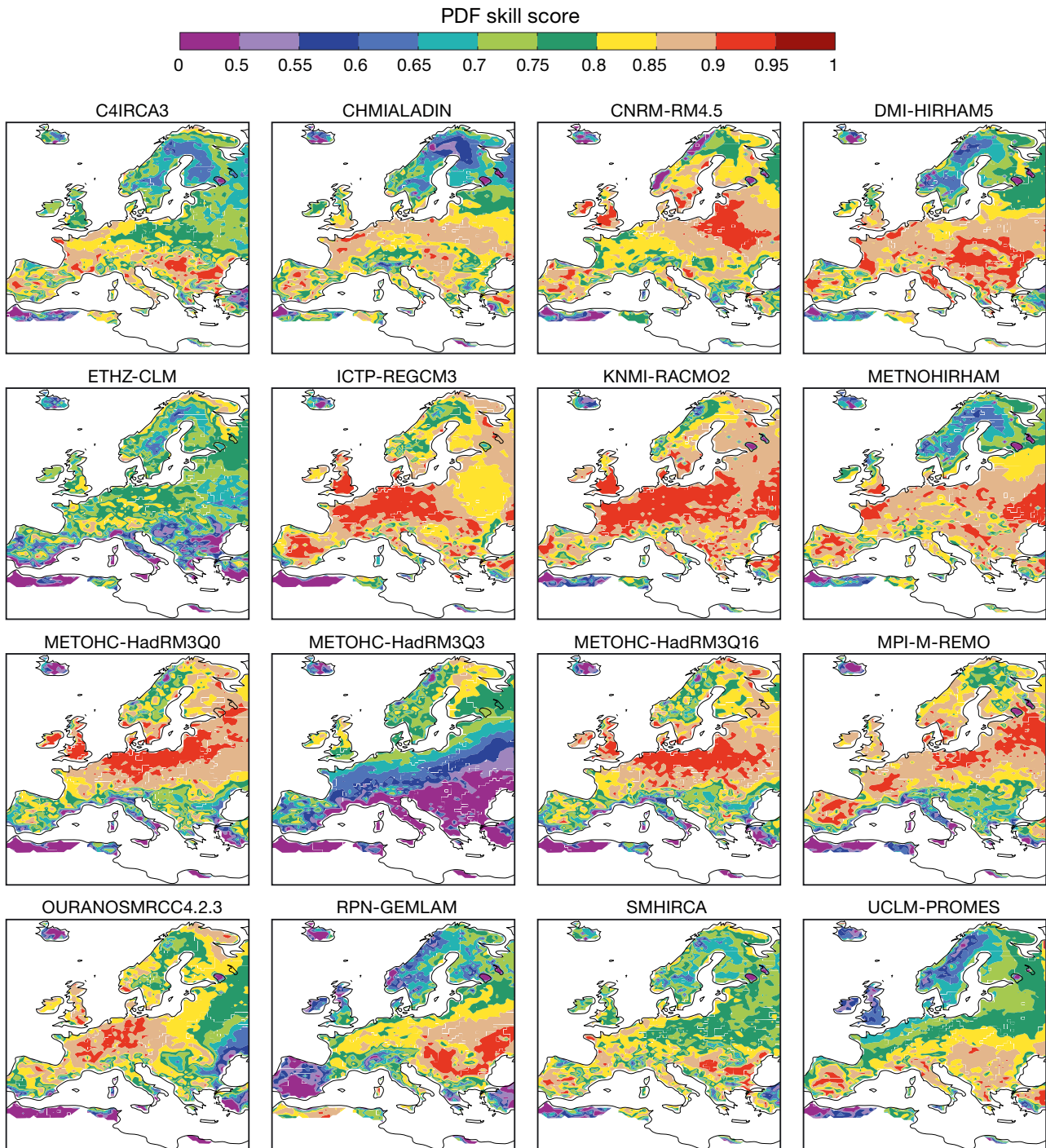
Fig. 4. Skill scores for daily maximum temperature during summer (JJA) from 1961–1990 for the 16 regional climate models (RCMs). The SS is dimensionless and takes a number between 0 and 1

tion in precipitation became worse with increasing precipitation in most models. One exception was MRC4.2.3, which simulated drier conditions than indicated by the observations. This overestimated heavy precipitation is in agreement with the findings of van Meijgard et al. (2008), who investigated precipitation extremes in a majority of the RCMs used here as compared to the E-OBS data. Here we note that this may be related to precipitation associated with extreme events in E-OBS being too low (Hofstra et al. 2009).

In summer the SSs were slightly lower than in winter in a majority of the models and in large parts of the
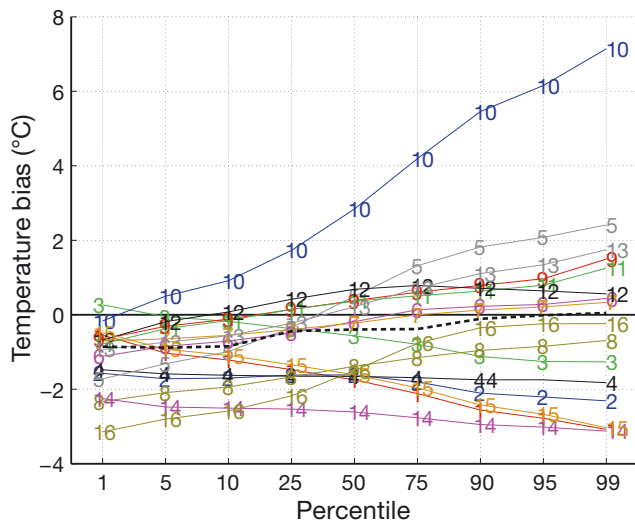
Fig. 5. For each model, the area-averaged bias over European land (EU) areas in summer (JJA) daily maximum temperature compared to E-OBS at different percentiles. Model names given in Table 1. Dashed black line: median of the 16 models

domain (not shown). This may be related to the fact that the distribution of daily precipitation is wider in summer than in winter, which makes it more difficult to achieve an equally good match, as we used the same bin width in both seasons. Alternatively, it may be an indication of the fact that the RCMs are more free to develop their own climate, as the influence of the large-scale circulation is weaker in summer than in winter. However, it should also be noted that some models were equally as good in summer as in winter. This holds for 2 of the 3 members of the Hadley Centre perturbed physics ensemble (HadRM3Q0 and HadRM 3Q16), while the third member (HadRM3Q3) showed considerably lower SSs. As in winter, there was a tendency for many models to simulate too much heavy precipitation (Fig. 7). Again this was most pronounced for the upper-end percentiles. In low and central parts of the probability distribution, the agreement was better, but with a systematic dry bias in all models at the lower quartile.

3.3. Monthly precipitation

Monthly statistics for wintertime precipitation showed that the models generally had the most problems in representing precipitation in the Mediterranean region, while SSs were relatively high in the British Isles (Fig. 8). The figure shows that different models were best in different regions; there was no clear overall 'winner' that outperformed the others. The same was true also for summer. Skill scores were

about equal in summer and winter in most regions apart from the Mediterranean area, where the agreement between models and observations on average was better in summer.

Biases, relative to E-OBS, of monthly precipitation at different percentiles are shown in Fig. 9. The spread between different RCMs was large, and for some models the biases were large, in particular at low percentiles. All models but CNRM gave too much precipitation in winter. In summer, large biases were seen at low percentiles, while biases were smaller than in winter at mid- and high percentiles. There was no clear dry or wet bias in summer as in winter, but there was a tendency for increasingly wet biases (or decreasingly small dry biases) at higher percentiles.

4. DISCUSSION

The methods used to calculate the SS metrics presented above are straightforward, and the SSs are seemingly objective. However, as there is a range of choices made in selecting data, calculating SSs and combining them into final weights, the exercise is of course associated with a large degree of subjectivity. In this section, we investigate how the results may change by using different data sets or other methods for area averaging in a region, for calculating the individual SS metrics and for combining the weights.

4.1. Sensitivity to the choice of bin width and threshold for precipitation data

For daily precipitation, we note that the results were sensitive to the choice of threshold that we set at 1 mm d^{-1} as our reference. By instead choosing a threshold of 0 mm d^{-1}, we obtained very high SSs, while lowering the threshold to a small number like 0.1 mm d^{-1} led to much lower SSs (not shown). This is related to the fact that many RCMs had too many days with drizzling rain with small (i.e. <1 mm d^{-1}) amounts of precipitation. If the threshold was 0 mm d^{-1}, these simulated drizzle events tended to fall in the same bin as the observations did, as no precipitation was registered by the collectors on dry days. If, however, we used a slightly higher threshold (e.g. 0.1 mm d^{-1}), the agreement was no longer there as the models simulated too many precipitation events in the first bin. With increasingly larger thresholds, approaching 1 mm d^{-1}, the SSs improved. Also, the bin width was important, but only if we used thresholds <1 mm d^{-1}, when a smaller bin width (e.g. 0.25 mm d^{-1}) led to lower SSs. The results indicate that the chosen threshold of 1 mm d^{-1}, which is a standard number used in climate impact related
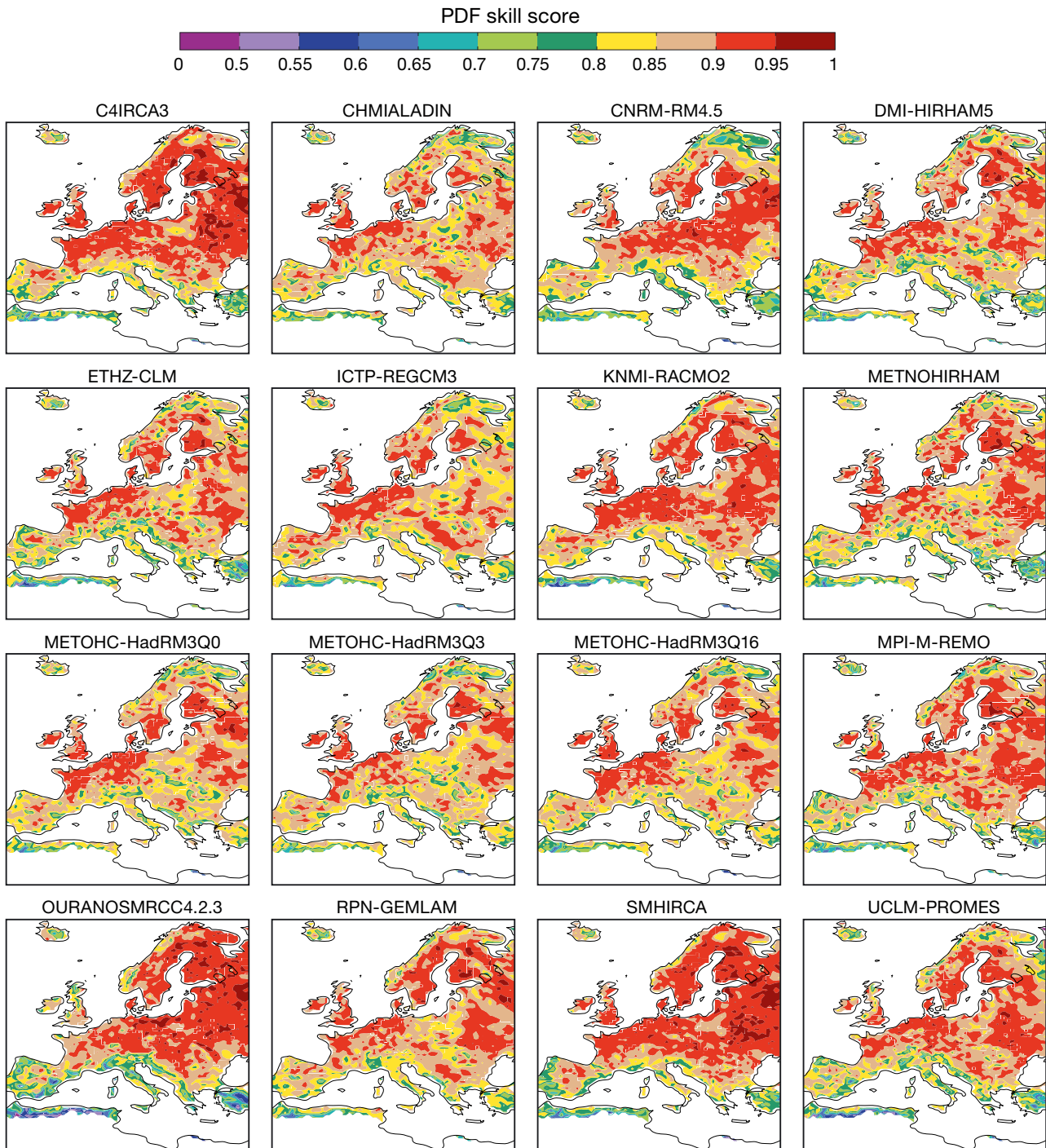
Fig. 6. Skill scores for daily precipitation (>1 mm d^{-1}) during winter (DJF) from 1961–1990 for the 16 regional climate models (RCMs). The skill score is dimensionless and takes a number between 0 and 1

research, appears to be a good threshold in that it removes drizzle discrepancies between observations and models. Finally, we note that low intensity precipitation does not contribute significantly to the total amount of precipitation (Dai 2001), further justifying a threshold of 1 mm d^{-1} for daily data.

4.2. Sensitivity to the choice of daily or monthly precipitation data

When comparing daily and monthly frequency distributions, it is important to notice that monthly values are the average of all days in a month including both
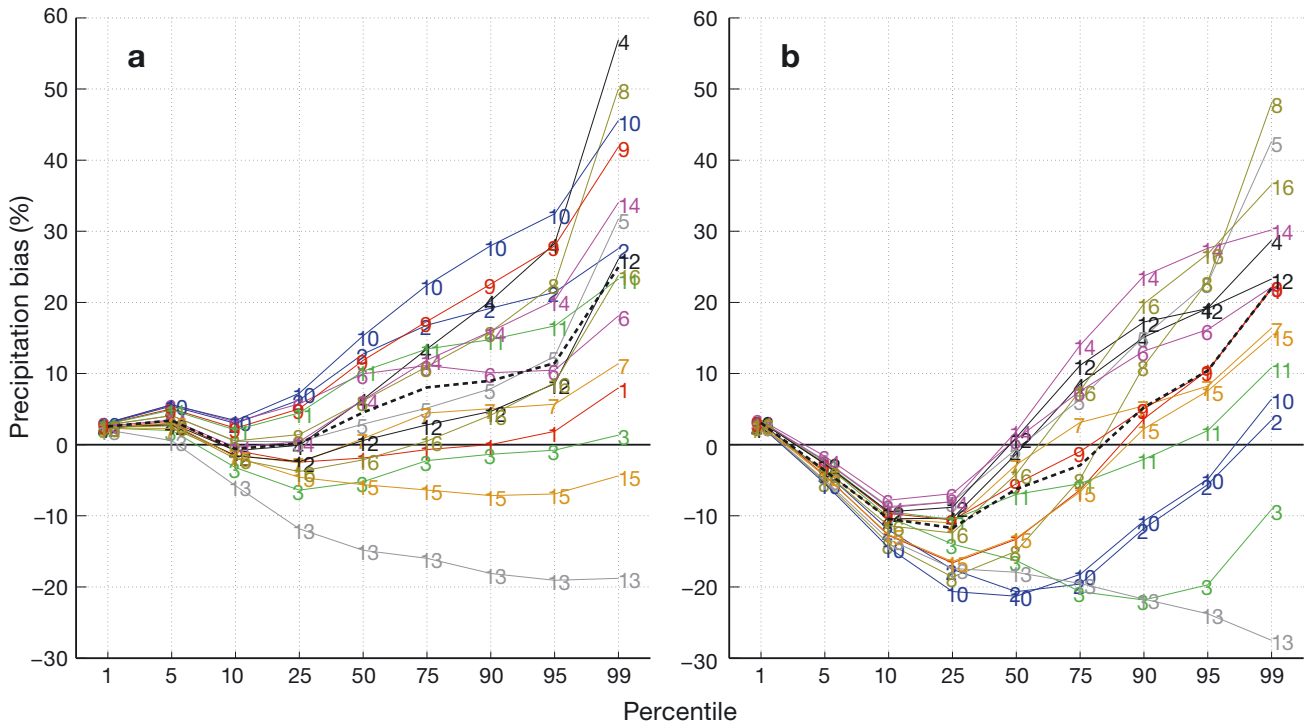
Fig. 7. For each model, the area-averaged bias over European land areas in comparison to E-OBS at different percentiles for (a) winter (DJF) and (b) summer (JJA). Model names given in Table 1. Dashed black line: median of the 16 models
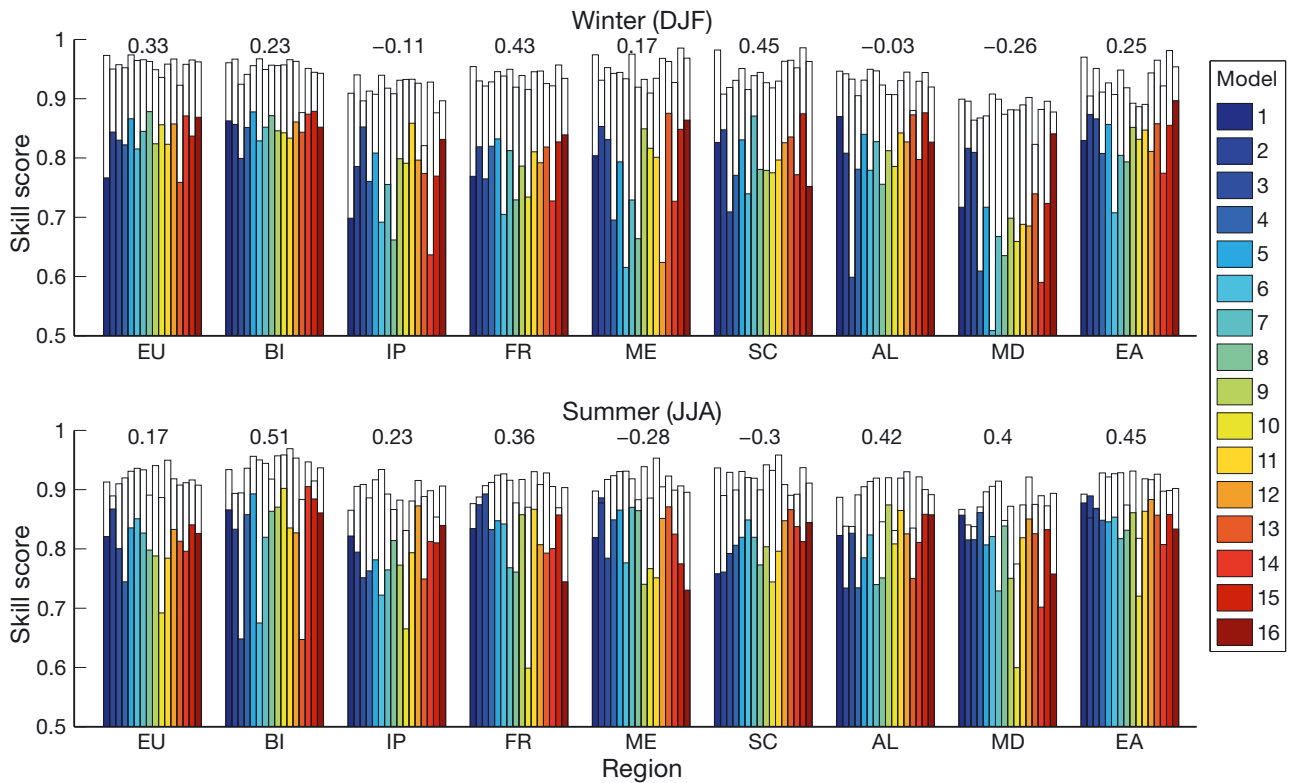

Fig. 8. Skill scores calculated with the method in Eq. (1) for precipitation, based on monthly (coloured bars) and daily (open bars) data for the European domain (EU) and the 8 sub-domains from Fig. 1. Numbers above each group: correlation coefficient between SSs based on daily and monthly statistics. Model numbers (1–16) in the colour key are the same as in Table 1. The SS is dimensionless and takes a number between 0 and 1
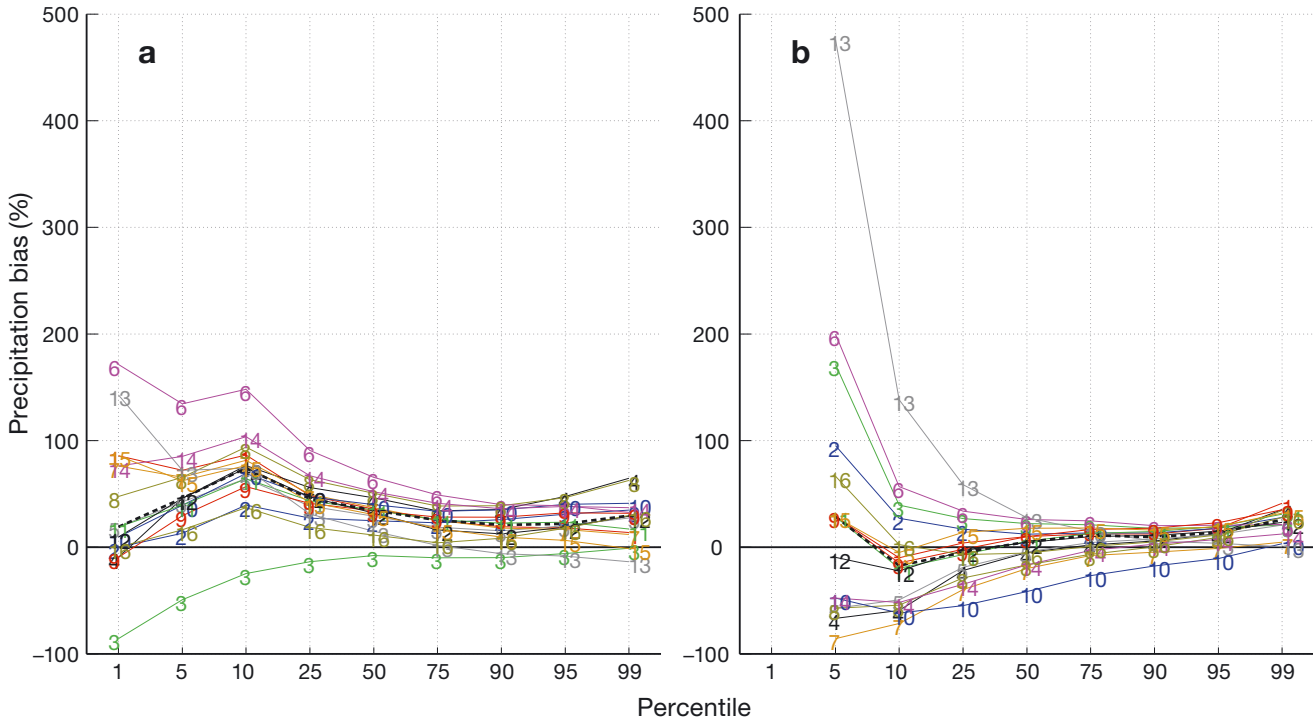
Fig. 9. For each model, the area-averaged bias over European land areas in monthly precipitation compared to E-OBS at different percentiles for (a) winter (DJF) and (b) summer (JJA). Model names given in Table 1. Dashed black line: median of the 16 models

dry days with no precipitation and days with light, moderate and/or extreme precipitation. This implies, on one hand, that monthly data include data from days not included in daily PDFs (as days with no precipitation fall below our chosen threshold, see Section 4.1). On the other hand, extreme daily precipitation amounts can be smoothed in accumulated monthly totals that also include a large number of days with little or no precipitation. In fact, it can be seen that the more the average is made towards annual values, the more Gaussian the distribution is (Lettenmaier 1995). An additional complication when comparing daily and monthly data is that the distribution of monthly data depends on the autocorrelation structure of the daily data. Thus, even if a model got the distribution of the daily data exactly right, this will not necessarily result in a correct monthly distribution.

Comparing Figs. 7 and 9 reveals large differences in the biases of monthly and daily data. While daily data, only including data above the threshold, showed relatively good agreement to observations at the low end of the PDF, monthly data including dry days disagreed the most at this part of the PDF. Differences in agreement with observations between daily and monthly data were also manifested in the calculated SSs. Fig. 8 shows the SSs based on both monthly and daily precipitation data calculated with the same metric (Eq. 1). As can be seen, the daily based SSs were higher than

those based on monthly data. The reason for this may be related to the low threshold for the daily data removing the dry days — for which models and observations tended to disagree (see Section 4.1) — from the statistics. We note that the ranking of the models differed significantly depending on which data set was used here. It was not the case that a 'good' model in representing daily data was necessarily 'good' in terms of representing monthly statistics. This can be seen by comparing the ranking of the models by looking at individual bars or by noting the low degree of correlation between the 2 data sets for each region. The correlation coefficient (0.51) was statistically significant at the 95% significance level only for the British Isles in summer. The differences in these 2 data sets clearly illustrate that monthly and daily precipitation statistics behave differently, such that both should be considered for model evaluation purposes.

4.3. Sensitivity to the sampling models for daily maximum/minimum

In addition to the choice of data sets, different ways of calculating maximum and minimum temperatures in the RCMs also introduces a source of uncertainty. The longer time steps for storing data in CLM and RM4.5 (Table 1) imply that those models may underestimate

extremes as they may not coincide with the fixed time steps every 3 or 6 h. Judging from the calculated SSs, it appears that CLM did not perform worse than the other models for minimum temperatures in winter (Fig. 2). RM4.5, on the other hand, showed relatively low SSs. A closer look at biases in different parts of the PDFs revealed that RM4.5 had exceedingly large warm biases closer toward the cold side of the PDF, both in Scandinavia (Fig. 3), but also in other areas (not shown). However, it is difficult to judge if this is indeed a problem with sampling, as RM4.5 and ALADIN behaved very similarly, both in terms of SSs and bias structure. The situation in summer showed that SSs for maximum temperatures were a bit lower in CLM than in some of the other models, including RM4.5 (Fig. 4). This may indicate that the absolute maximum was not captured by the 8 instantaneous numbers available for each day in CLM. However, we also note that CLM was warm biased in the warm part of the PDFs
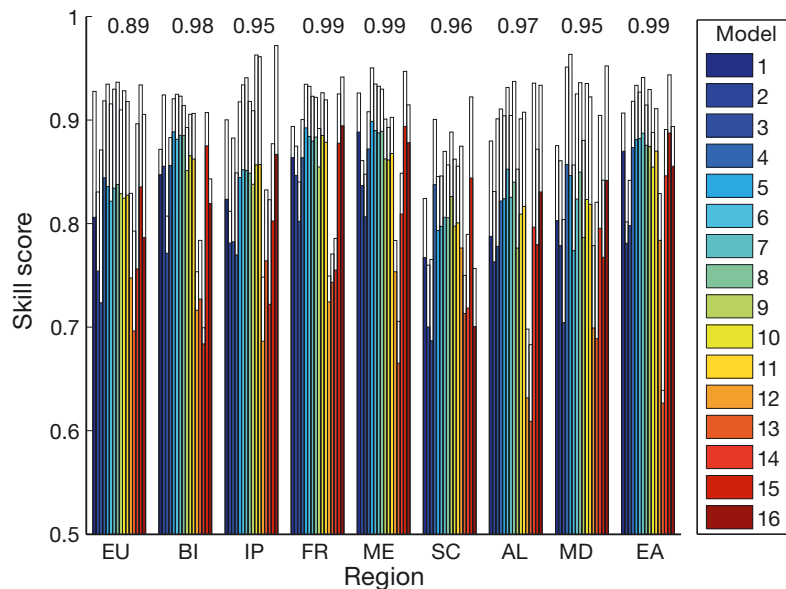

Fig. 10. Skill scores calculated for winter (DJF) daily minimum temperatures. Area averaging over the regions was done according to Method (3) (outlined in Section 2.5), by calculating SSs in each grid box individually (coloured bars) and according to Method (1) by pooling all individual grid box data into 1 common probability distribution (open bars) for the European domain (EU) and the 8 sub-domains from Fig. 1. Numbers above each group: correlation coefficient between SSs based on the 2 sets of area averages. Model numbers (1–16) in the colour key are the same as in Table 1. The SS is dimensionless and takes a number between 0 and 1

(above the 75th percentile, reaching about 2°C at the 95th and 99th percentiles, cf. Fig. 5), implying that even higher temperatures would not necessarily lead to a higher SS. RM4.5, which achieved a relatively high SS in summer, showed an increasingly cold bias at higher percentiles in summer, possibly indicating that sampling was too infrequent to record the highest temperatures in the course of the day. For the other RCMs, differences in time steps for calculating maximum and minimum temperatures were between 2 and 15 min, which may also have had some impact on the results. However, we assume that these differences were relatively small compared to other differences between the different RCMs in terms of how they represent sub-grid scale processes in their parameterisation schemes and also in terms of the relatively coarse horizontal resolution considered.

4.4. Sensitivity to the choice of area averaging

Area-averaged SSs for all RCMs for the 8 different sub-regions as calculated by Methods (1) and (3) outlined in Section 2.5 were similar as exemplified for daily minimum winter temperatures in Fig. 10. The 2 methods differ, as particularly (1) is 'forgiving' in a sense that it mixes spatial and temporal variability. Consequently, SSs were generally higher in (1) com-

pared to (3), which retains more of the geographical information from the RCM. The 2 methods showed that the ranking between the different models did not change significantly depending on which method was used. This can be seen from the correlation coefficients that correlated the 2 sets of SSs with each other. Method (2) is somewhere in between (1) and (3) in terms of how much of the geographical information is kept. Skill scores calculated by method (2) were very similar to those shown in Fig. 10, also in terms of ordering of models (not shown). The implication is that a model that is 'good' in the comparison to observational data is also 'good' if area averaging is done in another way.

4.5. Sensitivity to the choice of SS metric

For monthly precipitation, we present SSs for winter (DJF) based on the 2 methods outlined above (Eqs. 1 & 7) in Fig. 11. The difference in absolute numbers was very large, with the Perkins et al. (2007) method (Eq. 1) giving much higher numbers compared to that used by Sánchez et al. (2009) (Eq. 7). The difference in absolute numbers calculated by the 2 methods was much larger than when comparing SSs based on monthly and daily data calculated with the same SS (Eq. 1) as presented in Fig. 8. In addition to the inclusion of more days with little or no precipitation, as discussed above, the differ-
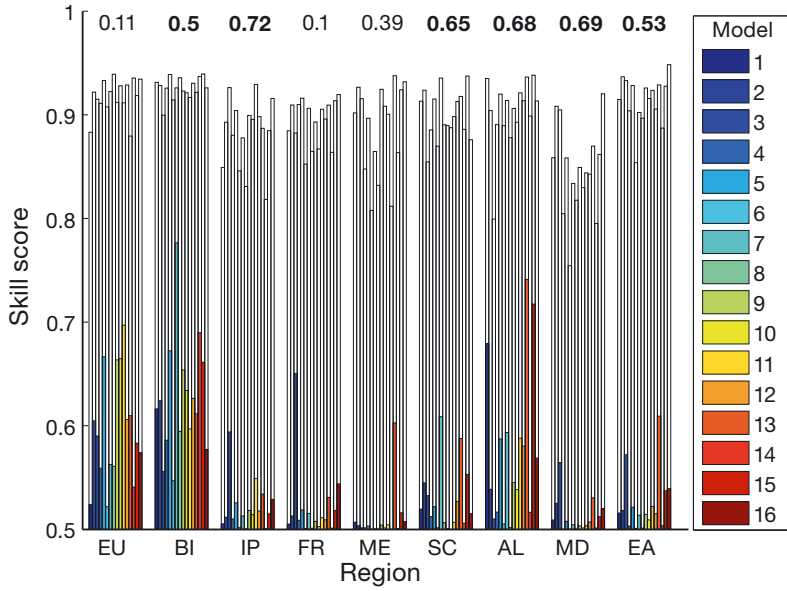
Fig. 11. Skill scores for winter (DJF) precipitation calculated with the method from Sánchez et al. (2009) (coloured bars) and according to Perkins et al. (2007) (open bars) data for the European domain (EU) and the 8 sub-domains from Fig. 1. Numbers above each group: correlation coefficient between SSs based on daily and monthly statistics. **Bold**: correlations are statistically significant. Model numbers (1–16) in the colour key are the same as in Table 1. The SS is dimensionless and takes a number between 0 and 1

which models were the best (or worst) in all regions, and the degree of correlation between the methods was always much lower compared to that for different methods of area averaging as described above. On the other hand, the degree of correlation between the 2 methods was statistically significant and higher in more areas than in the case of monthly and daily data as discussed above. A similar picture was also seen for other seasons (not shown).

4.6. Combined weights

The results for the different seasons, variables and areas generally showed that there are areas/seasons for which some models were better than others for the different variables (Table 2). In the individual seasons, it was not possible to pick a 'winner' in all areas, but taken as an average over the whole year, RACMO2 stands out with the highest ranking in 6 of the areas investigated as well as for the whole continent. This is also the model that ranked as the best in terms of overall performance. The second best in an overall sense was the high-sensitivity HadRM3Q16, which showed a good ranking in winter and summer while it was

ent formulations of SSs also contributed to the larger differences. Specifically, the Sánchez et al. (2009) method, in which 5 scores between 0 and 1 are multiplied, leads to overall lower SSs. We note here that the SSs calculated by the 2 methods did not agree on

Table 2. Ranking of the models (numbers as in Table 1) based on the calculated SSs for different regions, variables and seasons. 'Overall' shows the overall performance calculated as an average of the annual mean SSs for all regions. w_i: final weight for each model (see Section 2.7); Rank(w_i): ranking according to w_i; Rank(sum): ranking according to the sum of all ranks based on the individual SSs (Eq. 1) for each variable, region and season. 'Season' (DJF: winter, MAM: spring, JJA: summer, SON: autumn) and 'Region' (abbreviations as in Fig. 1) columns also show the ranking according to the sum of all ranks based on the individual SSs (Eq. 1) in each season and region, respectively

Model number	w_i	Overall Rank (w_i)	Rank (sum)	Season DJF	MAM	JJA	SON	Region EUR	BI	IP	FR	ME	SC	AL	MD	EA
1	0.908	9	9	11	9	7	8	10	11	8	10	9	10	4	4	7
2	0.906	10	7	6	10	12	5	8	6	10	4	6	14	11	2	9
3	0.910	7	13	5	13	15	11	12	12	7	15	15	9	14	10	11
4	0.903	11	14	8	12	13	15	14	16	13	11	14	16	12	7	12
5	0.912	5	5	14	7	3	1	4	7	9	6	4	4	7	11	4
6	0.885	14	4	2	2	4	16	2	2	4	2	5	13	2	9	8
7	0.932	1	1	1	1	5	6	1	5	3	1	1	1	1	1	1
8	0.910	6	10	7	4	14	13	3	8	12	12	10	12	9	8	10
9	0.913	4	3	4	6	6	10	5	4	2	5	2	6	8	6	3
10	0.896	13	12	10	11	10	14	13	3	11	14	12	7	16	14	15
11	0.917	2	2	3	8	2	9	6	1	1	3	3	3	6	3	5
12	0.901	12	11	12	15	1	4	11	9	5	8	13	5	10	12	13
13	0.878	15	15	9	14	16	7	16	15	15	13	11	11	15	15	14
14	0.869	16	16	16	16	9	12	15	13	16	16	16	15	13	16	16
15	0.913	3	6	13	5	8	3	7	10	14	7	7	2	5	13	2
16	0.909	8	8	15	3	11	2	9	14	6	9	8	8	3	5	6

intermediate among the other models during the transition seasons. A few models that performed less well than the others in most areas and seasons can be identified from Table 2. Many of the other models showed good performance during part of the year, or in part of the domain, and worse agreement in others. An example of such behaviour is REMO, which was the overall best model in summer, while its performance during spring ranked among the poorest models. The latter may be related to the temperature biases close to 0°C, as discussed above. Other similar examples include the RCA3.0 and ALADIN, whose performance was among the best in some areas, while in others those particular models ranked among the worst. These results clearly show that depending on area and season of interest, different models have different skills in representing the observed climate. An implication of this is that a general measure of overall skill in the whole model domain, and/or for all seasons, is not necessarily the best measure of skill in individual regions and seasons.

4.7. Evaluation of the weighting

As a test of the use of the weights in order to improve an RCM ensemble, we tested whether a weighted model mean outperforms an unweighted arithmetic mean. The weighted ensemble mean (WEM$_j$) was calculated as

$$\mathrm{WEM}_j = \frac{\sum\limits_{i=1,16} (w_{irs} \cdot \mathrm{VAR}_j)}{\sum\limits_{i=1,16} w_{irs}} \qquad (9)$$

where w_{irs} is the model specific weight (Eq. 8) and VAR$_j$ is the variable in question. Weighted ensemble mean seasonal averages showed very small differences with regard to the unweighted averages (not shown). For daily maximum and minimum temperature, this means generally <0.1°C for most grid boxes and correspondingly <±30% for precipitation, although larger relative biases do occur in mountainous areas and in the dry Mediterranean area in summer. These small differences are a result of the small spread in the overall weights (Table 2).

An additional test was done in which we constructed more discriminating weights based on the overall ranking of the models (Table 2). This was achieved simply by assigning each model a number from 1 (the poorest model) to 16 (the best model) and then normalising so that their sum was 1. In this way, the spread between good and poor models was more pronounced than with the original weights, and the results showed slightly larger differences between the weighted and unweighted ensemble means. For daily maximum and minimum temperature, differences between the 2 means were up to 0.3 to 0.4°C for some parts of the domain (not shown). Even if these differences between the 2 ensemble means were relatively small, the 2 did differ significantly in a statistical sense in a relatively large fraction of the model domain. This is presented in Table 3, which also shows to what extent the weighting leads to an overall improvement or worsening of the agreement between the ensemble mean and the observations. It can be seen that the weighting leads to a statistically significant improvement in a larger fraction of the domain than where it leads to a deterioration. This holds true for all variables and seasons with the exceptions of maximum temperature and precipitation in summer and minimum temperature in winter. Similarly, for the weighted ensemble mean calculated by Eq. (9), there were improvements in larger fractions of the area compared to where the results get worse (not shown). These results show that weighting mostly leads to small, but statistically significant, improvements compared to unweighted ensemble means. However, the results also underline the fact that the weights are based on the overall ranking that takes into account several variables and seasons implying that weighting may lead to a worse agreement with observations.

Table 3. Areal fraction of the analysed domain in which the weighted and unweighted means differed, and fractions where weighting led to an improvement or a deterioration. Unit: %. DJF: winter, MAM: spring, JJA: summer, SON: autumn

Variable	Season	Differences	Improvement	Deterioration
Daily max. temperature	DJF	97.0	69.5	27.5
	MAM	86.7	57.6	29.1
	JJA	81.4	30.2	51.2
	SON	97.9	84.3	13.6
	ANN	93.7	66.1	27.6
Daily min. temperature	DJF	90.0	25.7	64.3
	MAM	79.4	47.5	31.9
	JJA	96.7	80.8	16.0
	SON	91.1	53.4	37.7
	ANN	89.0	57.6	31.3
Daily precipitation	DJF	82.6	60.9	21.7
	MAM	70.1	48.9	21.2
	JJA	76.7	36.0	40.6
	SON	66.0	37.8	28.2
	ANN	78.0	44.3	33.7

5. SUMMARY AND CONCLUSIONS

• Biases in daily minimum temperature were large in northern Europe during winter, most notably for low end percentiles, when most models tended to simulate too warm conditions. These differences between models and observations tended to be largest in northern Scandinavia. Possibly, this discrepancy is more a consequence of the fact that observations and models represent different features than an indication of a systematic model error in this area.

• The highest SSs for daily maximum temperature in summer were seen in parts of central Europe in many models. Worse agreement was obtained in Scandinavia and also in southern Europe in some models. Biases in different regions were different, both in sign and amplitude, in different parts of the probability distributions among the RCMs.

• Biases in daily precipitation were most pronounced in the wettest part of the probability distribution where the RCMs tended to overestimate precipitation compared to the E-OBS data set. The overestimation grew with increasing amounts of precipitation. Skill scores were higher for precipitation than for temperature due to a good correspondence between models and observations for moderate precipitation.

• The calculated SSs and the evaluation performed here indicated that some models performed poorly for some variables and seasons. The underlying reasons for this were not revealed by our analysis, although results of this type may assist in identifying potential problems in different models.

• Applied on the present data set, the Perkins et al. (2007) method (Eq. 1) for calculating SSs gave a relatively small spread among the models. This is not surprising, as the method, by definition, integrates over the entire probability distributions. Further, positive and negative biases gave similar SSs, and biases in different parts of the probability distribution had equal influence on the resulting SS.

• The evaluation performed here over different variables, seasons and regions showed that some models can be better/worse than the others in an overall sense but that no model is best/worst in all aspects. The evaluation showed that some models performed well in some regions and seasons and poorly in others. An implication of this is that preferably the whole ensemble should be used in studies of climate change and impact studies. Another implication is that weighting of a model ensemble possibly should not be based on overall performance measures of the models, but instead should be based on different weights for different regions, seasons, variables, etc., depending on individual applications.

• The sensitivity to area averaging over the regions in Europe was small as the resulting SSs and ranking of models were similar regardless of whether area averaging was done before or after calculating SSs.

• The sensitivity to choice of SS metric was very large. The 2 different methods used for calculating SSs for monthly precipitation gave substantially different numbers and also different rankings of the models.

• The sensitivity to choice of input data was large. The same SS metric based on daily data excluding dry events, or monthly data including both wet and dry events, gave very different results in terms of model ranking. This was most likely a result of the fact that the underlying PDFs differed, and that dry days were excluded from the daily data.

• We found that weighted ensemble means were closer to observations than corresponding unweighted ensemble means for most, but not all, variables and seasons. This is the result of there being statistically significant improvements in a larger fraction of the domain than where the agreement deteriorated. This held true both for the original weights calculated from the SSs presented here and for more discriminating weights based on a ranking of the models. The differences, however, were small: for daily minimum and maximum temperatures, generally below 0.1°C for most grid boxes for the original weights, and most often below 0.3 to 0.4°C for the more discriminating weights.

LITERATURE CITED

Beniston M, Stephenson DB, Christensen OB, Ferro CAT and others (2007) Future extreme events in European climate: an exploration of regional climate model projections. Clim Change 81(Suppl 1):71–95

Boberg F, Berg P, Thejll P, Gutowski WJ, Christensen JH (2009a) Improved confidence in climate change projections of precipitation evaluated using daily statistics from the PRUDENCE ensemble. Clim Dyn 32:1097–1106

Boberg F, Berg P, Thejll P, Gutowski WJ, Christensen JH (2009b) Improved confidence in climate change projections of precipitation further evaluated using daily statistics from ENSEMBLES models. Clim Dyn doi:10.1007/s00382-009-0683-8

Böhm U, Kücken M, Ahrens W, Block A, Hauffe D, Keuler K, Rockel B, Will A (2006) CLM — the climate version of LM: brief description and long-term applications. COSMO Newsletter 6

Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of the century. Clim Change 81:7–30

Christensen OB, Drews M, Christensen JH, Dethloff K, Ketelsen K, Hebestadt I, Rinke A (2006) The HIRHAM Regional Climate Model Version 5 (β). DMI, Tech Rep 06-17

Christensen JH, Hewitson B, Busuioc A, Chen A and others (2007) Regional climate projections. In: Solomon S, Qin D, Manning M, Chen Z and others (eds) Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, p 847–940

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M (2010) Weight assignment in regional climate models. Clim Res 44:179–194

Collins M, Booth BB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2010) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim Dyn doi: 10.1007/s00382-010-0808-0

Côté J, Gravel S, Methot A, Patoine A, Roch M, Staniforth A (1998) The operational CMC/MRB Global Environmental Multiscale (GEM) model. I. design considerations and formulation. Mon Weather Rev 126:1373–1395

Dai A (2001) Global precipitation and thunderstorm frequencies. I. Seasonal and interannual variations. J Clim 14:1092–1111

Déqué M, Somot S (2010) Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments. Clim Res 44:195–209

Farda A, Déqué M, Somot S, Horanyi A, Spiridonov V, Toth H (2010) Model ALADIN as regional climate model for central and eastern Europe. Stud Geophys Geod 54:313–332

Giorgi F, Mearns LO (1999) Introduction to special section: regional climate modeling revisited. J Geophys Res 104:6335–6352

Haugen JE, Haakenstad H (2006) Validation of HIRHAM version 2 with 50 and 25 km resolution. RegClim Gen Tech Rep 9:159–173

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded dataset of surface temperature and precipitation. J Geophys Res 113:D20119 doi:10/1029/2009JD011799

Hewitt C, Griggs D (2005) Ensembles-based predictions of climate changes and their impacts. Eos Trans Am Geophys Union 85:566

Hofstra N, Haylock M, New M, Jones PD (2009) Testing E-OBS European high-resolution gridded dataset of daily precipitation and surface temperature. J Geophys Res 114:D21101 doi:10.1029/2009JD011799

Jacob D (2001) A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. Meteorol Atmos Phys 77:61–73

Jacob D, Bärring L, Christensen OB, Christensen JH and others (2007) An inter-comparison of regional climate models for Europe: design of the experiments and model performance. Clim Change 81:31–52

Kjellström E, Bärring L, Gollvik S, Hansson U and others (2005) A 140-year simulation of European climate with the new version of the Rossby Centre regional atmospheric climate model (RCA3). Rep Meteorol Climatol 108. Swedish Meteorological and Hydrological Institute, Norrköping

Kjellström E, Bärring L, Jacob D, Jones R, Lenderink G, Schär C (2007) Modelling daily temperature extremes: recent climate and future changes over Europe. Clim Change 81:249–265

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R and others (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int J Climatol 22:1441–1453

Klok EJ, Klein Tank AM (2009) Updated and extended European dataset of daily climate observations. Int J Climatol 29:1182–1191

Lettenmaier D (1995) Stochastic modeling of precipitation with applications to climate model downscaling. In: Von Storch H, Navarra A (eds) Analysis of climate variability; applications of statistical techniques. Springer, Berlin, p 197–212

Moberg A, Jones PD (2004) Regional climate model simulations of daily maximum and minimum near-surface temperatures across Europe compared with observed station data 1961–1990. Clim Dyn 23:695–715

Nikulin G, Kjellström E, Hansson U, Jones C, Strandberg G, Ullerstig A (2010) Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. Tellus Ser A Dyn Meterol Oceanogr (in press) doi:10.1111/j.1600-0870.2010.00466.x

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. J Clim 20:4356–4376

Plummer D, Caya D, Coté H, Frigon A and others (2006) Climate and climate change over North America as simulated by the Canadian regional climate model. J Clim 19:3112–3132

Radu R, Déqué M, Somot S (2008) Spectral nudging in a spectral regional climate model. Tellus Ser A Dyn Meterol Oceanogr 60:898–910

Räisänen J, Hansson U, Ullerstig A, Döscher R and others (2004) European climate in the late 21st century: regional simulations with two driving global models and two forcing scenarios. Clim Dyn 22:13–31

Rockel B, Woth K (2007) Extremes of near-surface wind speed over Europe and their future changes as estimated from an ensemble of RCM simulations. Clim Change 81:267–280

Sánchez E, Gallardo C, Gaertner MA, Arribas A, Castro M (2004) Future climate extreme events in the Mediterranean simulated by a regional climate model: a first approach. Glob Planet Change 44:163–180

Sánchez E, Romera R, Gaertner MA, Gallardo C, Castro M (2009) A weighting proposal for an ensemble of regional climate models over Europe driven by 1961–2000 ERA40 based on monthly precipitation probability density functions. Atmos Sci Lett 10:241–248

Uppala SM, Kållberg PW, Simmons AJ, Andrae U and others (2005) The ERA-40 re-analysis. QJR Meteorol Soc 131:2961–3012

van der Linden P, Mitchell JFB (eds) (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office, Hadley Centre, Exeter

van Meijgaard E, van Ulft LH, van de Berg WJ, Bosveld FC, van den Hurk BJJM, Lenderink G, Siebesma AP (2008) The KNMI regional atmospheric climate model RACMO, version 2.1. TR-302, KNMI, De Bilt

Zadra A, Caya D, Côté J, Dugas B, and others (2008) The next Canadian Regional Climate Model. Phys Can 64:75–83